Research Journal of Applied Sciences, Engineering and Technology 5(14): 3824-3829, 2013
DOI:10.19026/rjaset.5.4531
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: October 17, 2012 Accepted: January 03, 2013 Published: April 20, 2013

Research Article

The Design and Implementation of Intrusion Detection System based on Data Mining Technology

Qinglei Zhou and Yilin Zhao
School of Information Engineering, Zhengzhou University, Zhengzhou 450001, China

Abstract: Intrusion detection technology is a research hotspot in the field of information security. This study introduces the types of traditional intrusion detection and data mining technology; Aiming at the defects and limitations of current intrusion detection system, the study has fused the data mining technology into intrusion detection model, and has designed and implemented the intrusion detection system based on data mining technology with the preliminary research and exploration.

Keywords: Data mining technology, information security, intrusion detection

INTRODUCTION

With the rapid development of the internet, the various attacks, which emerge endlessly in the network, have become a major threat to network and information security. Traditionally, network users usually use firewall as the first line of defense for security. But with attacking tools and means becoming much more complicated, simple firewall is difficult to resist various attacks; therefore people put forward a kind of technology which can discover in time and report unauthorized or abnormal phenomena in the system, named intrusion detection technology (Jiang *et al*., 2000). For existing complex attack behaviors in the network, it is an important study issue how to establish an effective intrusion detection model (Su and Fu, 2007) by network security experts. This study makes a preliminary in this field of research and exploration and implements intrusion detection model system based on data mining technology (Lu *et al*., 2003).

RELATIONAL TRADITIONAL INTRUSION DETECTION SYSTEM

There are two types of traditional intrusion detection system (Gu and Sun, 2006).

Anomaly detection: The aim is to detect abnormal behavior for host or network. To summarize the characteristics of normal operation, it is considered to be invaded when there is significant deviation between the user activities and the model of the normal behavior. The detection model should have low missing alarm rate and high misinformation rate. At the same time, the intrusion detection system needs to constantly update model library to keep up the development of the invasion technology.

Misuse detection: The task is to detect the matching degree with known unacceptable behavior. The system acquires the characteristics of abnormal operation behavior, and then establishes characteristics model library of aggressive behavior. When the behavior of user or system matches the library of records, the system regards it as invasion behavior. As to attacks that have found, this detection method can accurately report the attack type in detail; but for a new attack, the effect is limited, characteristic mode library must be continually updated.

The effectiveness, adaptability and extensibility are important indexes to evaluate the quality of the intrusion detection system. The current intrusion detection systems usually use statistics analysis method to analysis the known intrusion method and system vulnerability. The rules which are designed for the specific system environment and detection method are provided by security experts through the manual coding. It makes system inefficient and limits the ability of scalability and self-adaption of the system.

Aiming at the defects of current intrusion detection system, this study views data as the center point, applies data mining technology into intrusion detection system, designs and implements intrusion detection system model based on data mining technology and researches the performance of the system.

DATA MINING TECHNOLOGY

In order to overcome the limitations of traditional intrusion detection system, a systematic and higher automation method should be employed in the design of intrusion detection system. Data mining is the kind of effective method. Data mining concentrates on analyzing from a lot of noise, fuzzy and random data

Corresponding Author: Yilin Zhao, School of Information Engineering, Zhengzhou University, Zhengzhou 45001, China

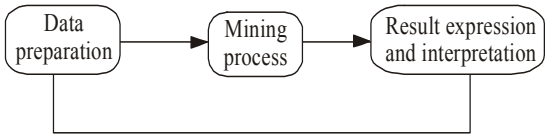

Fig. 1: The whole process of data mining

and extracts meaningful, potential useful information and knowledge.

Data mining process: The whole process of data mining is a repeated execution of the following three steps (Fig. 1):

- Collect the data need for mining
- Do the mining operation for the data
- Get the corresponding result and then express the result with a certain way

If the mining results are not satisfied with what we want to, the three steps will continue to be carried out until mining results are corresponding with our intention results.

Main advantages: The intrusion detection technology based on data mining can automatically extract the knowledge and model from the training data, accurately and effectively detect the actual intrusion and normal behavior mode, and reduce people's participation and the burden of the intrusion detection analyst, without the need of manual analysis and code intrusion pattern.

The intrusion detection system based on data mining technology has followed main advantages.

- **Adaptive ability:** According to the existing attack, experts analyze and create their characteristic model as the rule base of traditional intrusion detection system. If a kind of attack step over a longer period of time, so the original intrusion detection system rule base is difficult to get update timely. Because anomaly detection and signal matching mode is different in the application of data mining technology, which does not detect each signal, so new attack can be effectively detected.
- **Low false alarm rate:** The detection principle of the existing systems mainly relies on the simple signal matching. This stiff way makes the fact that report rate does not agree with actual situation. With the combination of data mining technology and intrusion detection technology, the system finds out regular pattern implicitly and filters out abnormal behavior signals; thereby it will reduce the false alarm rate of the system.
- **Strong intellectuality:** The system which applied data mining into intrusion detection not only can automatically extract undetectable behavior modes from a large number of network data, but also increase the accuracy of detection system.
- **High efficiency detection:** facing to mass data flow in the current network environment, the data should be detected by intrusion detection system, which has to be processed after pretreatment. The intrusion detection system can automatically preprocess data, extract useful data, and improve the detection efficiency.

INTRUSION DETECTION ALGORITHM BASED ON DATA MINING

Finding model is the core idea of data mining, which is applied into intrusion detection system through the data mining algorithm. The system can find the knowledge and rule which can be contributed to detect attack from the system log. Association rules mining algorithm (Gao and Wang, 2003) and clustering analysis algorithm have been applied into the intrusion detection system in this study.

Association rules mining: The detection system should have low false alarm rate and high misinformation rate. At the same time, the intrusion detection system needs to constantly update model library to keep up the development of the invasion technology. The aim of correlation analysis is to mine the correlation relationship between the different attributes of data and find out the rules to meet multiple attributes with certain support (support) and confidence degrees (confidence) constraint in the database. It usually uses association rules to express this kind of relation. In the intrusion detection, the system finds association between the attributes of the invasion audit data through the association rule algorithm and forms the judgment rules of the invasion.

M = {M1, M2,…Mk} is a set of all the attributes of behavior, S = {S1,S2,…Si} is a set of all behaviors. Si is a behavior of containing subset in M attribute set, Si $\subseteq$ M. S contains W, which is a subset of M, W$\subseteq$ S. Association rule is a form such as X=>Y expression:"X=>Y, the support degree is m% and the confidence degree is d%". X and Y item are attribute subsets of containing the same behavior, X$\cap$Y = $\emptyset$. The support degree S(support) and confidence degree C(confidence) are two rules of interest measurement in the rule. Support degree S refers to the probability of the behavior of containing X$\cup$Y in S set, namely P (X$\cup$Y). Confidence C refers to probability of the behavior of containing both X and Y subset in S set, namely the conditional probability P (Y|X). Such associations' rules can be expressed as the following form. S, C are rules support and confidence.

$$\text{Support}(X=>Y) = P(X \cup Y) \quad (1)$$

$$\text{Confidence}(X=>Y) = P(Y \mid X) \quad (2)$$

The support degree is vital importance of the measure to occasional rules; therefore, support degree is usually used to delete those rules which are not interesting. Confidence is to measure the accuracy of association rules. Correlation analysis generates association rules between the data item sets, and ensures that the degree of support and confidence are greater than the user reassigned minimum support degree (MIN_SUP) and minimum confidence (MIN_CONF), also known as the support degree threshold and confidence threshold value. This study uses the association rule mining algorithm called Apriori algorithm (Ning and Guo, 2006).

Association rules mining algorithm:

Input: database D, minimum support degree min_sup.

Output: all the frequent itemsets U in D.

Flow:

L_1= find_freguent_1-itemsets (D);
for (k=2;$L_{k-1}\neq\emptyset$;k++)
{
Ck = apriori_gen (Lk-1); //get the candidate subsets
for each transaction t∈D
{
Ct = subset(C_k, t);
For each candidate c∈Ct
c.count++;
}
Lk = {c∈C_k|c.count >= min_sup}
}
return L = $Y_k L_k$;
Procedure apriori_gen(L_{k-1})
For each itemset $l_1 L_{k-1}$
For each itemset $l_2 L_{k-1}$
If(l_1[1] = l_2 [1])∧(l_1[2] = l_2 [2])∧…∧ (l_1[k-2] = l_2 [k-2])∧(l_1[k-1] = l_2 [k-1])
{
c = $l_1 \infty l_2$;
If has_infrequen_subset (c, L_{k-1}) then
delete c;
else
add c to C_k;
}
return C_k;
}
Produrehas_infrequent_subset (c: candidate k itemset; L_{k-1}:frequent (k-1) itemset)
For each (k-1) subset s of c
If s is not"∈" L_{k-1}, then
return true
return false;

Generate association rules:

- For each frequent item sets l, produce all non vacuous subset for l.
- For each not vacuous subset sin l, if (support (l))/(support(s))>=min_conf, the output rule: "s=>(l-s)".
  The min_conf is minimum confidence threshold value.

Clustering analysis algorithm: Clustering analysis (Chen, 2010) is the process of dividing concentration objects into several similar kinds of objects according to the similarity features. Through the clustering algorithm, it makes the same objects of a class have high similarity, and inhomogeneous objects have low similarity. Its characteristic is that the divided data set or class is unknown. This study uses K-Means algorithm (Xu and Cai, 2008) as the clustering analysis algorithm.

K-Means algorithm:

- Assign initial Values for means ml, m2... mk
- Repeat assign each item to the cluster which has the encloses center
- Calculate new center for each cluster
- Until convergence criteria is met

The intrusion detection system using K-Means algorithm has followed advantages: simple algorithms, small computational complexity, satisfy the requirements of real time intrusion detection and easy to implement. First, this system presets a clustering radius. Second, use the first packet as the first clustering center. Third, calculate similarity between other packets and the entire clustering centers. If the similarity is less than or equal the mean value, the packet is divided into the corresponding clustering. And then recalculate mean value of the clustering center. But if the similarity is more than the mean value, the packet is indentified as a new clustering center.

THE DESIGN AND IMPLEMENT OF INTRUSION DETECTION SYSTEM BASED ON DATA MINING TECHNOLOGY

The structure of intrusion detection system based on data mining technology: This study designs and implements the intrusion detection system based on data mining technology as shown in Fig. 2.

System module function summary:

- **Sniffer:** Mainly acquire data, grab packets from network
- **Decoder:** Mainly decode and analyze the datagram, store the results
- **Preprocessor:** Transform the packet to the format for data mining, restructure and process code conversion before matching

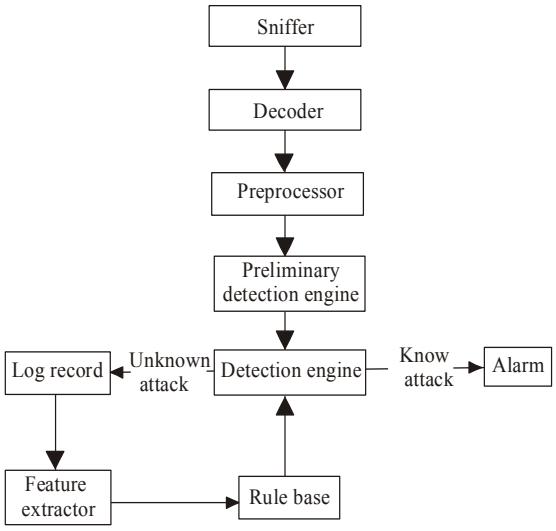

Fig. 2: Intrusion detection system structure diagram

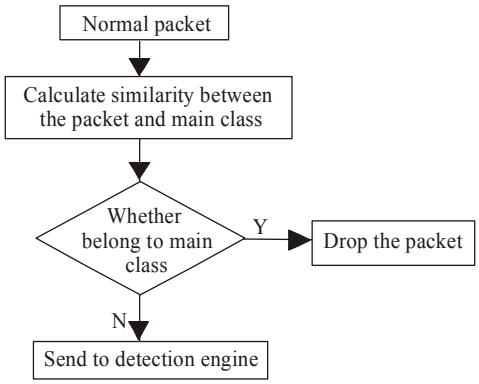

Fig. 3: The module workflow

- **Preliminary detection engine:** Mainly filter out normal network packets
- **Detection engine:** Mainly match rule. It uses K_Means algorithm as the clustering analysis algorithm.
- **Log records:** Include packets information which produced by unknown network normal behavior and unknown intrusion behavior.
- **Rule base:** Save some existing or new intrusion detection rules.
- **Feature extractor:** Make correlation analysis of the data in a log, conclude the new association rule, and add it to the rule base. It uses Apriori algorithm correlation analysis.
- **Alarm:** Transmit an alert when there is an abnormal behavior.

Workflow: The workflow of the intrusion detection system based on data mining is introduced as follows. Firstly, the sniffer grabs network packets which are analyzed by the decoder. Then preprocessor will process the parsing packets by calling pretreatment function. Secondly, after through the preliminary detection engine, normal packets will be discarded off, and the abnormal packets will be processed by detection engine. Through matching rule, it shows that there are invaded behaviors when successful. At the same time, the system will transmit an alert and prevent intrusion behavior. If it is not successful, the new network normal behavior model will be recorded into log. Finally, the system will make the correlation analysis for the log through the data mining algorithm. If there is a new rule generation, it will be added to the rule base.

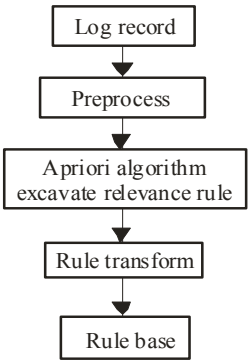

Fig. 4: The module workflow

Feature extractor: The workflow of preliminary detection engine using K_Means clustering analysis algorithm is shown in Fig. 3.

Feature extractor: The aim of feature extractor is to mine association rules through association rules mining algorithm. First it analyzes the abnormal packets, which had been processed by the pretreatment; and then obtains potential or new intrusion behavior patterns through the Apriori association rules algorithm and produce s the corresponding association rule set; Finally it transforms the rule into the intrusion detection rule and adds it to the rule base. The module workflow is shown in Fig. 4.

EXPERIMENTS

Experimental environments: This system uses the VC as the development tool, windows XP operating system, the Snort v2.4.3, libpcap v0.8.3 as caught tools and simulation network attack software IDS Informer v4.0. In this study, the intrusion detection system is developed by C language. In the rule matching aspects, we maintain the working principle of Snort. The experiment, which is representative, analyzes the relationship between attack mode database size and matching time.

Table 1: Description of attributes

Attribute name	Description	Category
Ip-len	Length of packet	Continuous type
Ip-ttl	Lifetime	Continuous type
Tcp-win	TCP window size	Continuous type
Ip-options_len	Length of the IP protocol rule option	Continuous type
Tcp-options-len	Length of the protocol rule option	Continuous type
Dsize	Payload size of packet	Continuous type
Udp-len	Length of UDP data packet	Continuous type

Table 2: Data sheet in example.txt

Attribute name	Average value	Mean absolute error
Ip-len	28090	15365
Ip-tl	80	24
Tcp-win	45862	21325
Ip-options-len	0.00142	0.00279
Tcp-options-len	8.16	7.968
Dsize	568.3	596.65
Udp-len	38956	13598

Table 3: Data clustering table

Cluster radius	Network normal behavior pattern class	Network abnormal behavior pattern class
1	0	142
2	1	136
3	3	128
4	6	112
5	9	91

Table 4: Data false detection rate table

Cluster radius	1	2	3	4	5	10
False detection rate (%)	0	0.06	0.15	0.36	0.71	0.86

Table 5: Data clustering table

Threshold	Network normal behavior pattern class	Network abnormal behavior pattern class
100	69	61
200	52	69
300	39	76
500	21	99
800	9	113

Table 6: Data false detection rate table

Threshold	100	200	300	500	800
False detection rate (%)	0.043	0.026	0.015	0.09	0.03

Experimental method: Through the libpcap grabbing random network packets, the network packets contain a large number of attributes, such as time (timestamp), the source host (src_host), the source port (src_port), service (service) etc. Many attributes are not big significance on cluster analysis, so some attributes which we do not need are discarded. Selected attributes are shown in Table 1.

The attribute data of packets are saved in example.txt. The sample data are shown in Table 2.

Experimental result: The aim is to research the performance of the intrusion detection system in this study. So when selecting different cluster radius and thresholds, what effect will have on the results?

- **The case of threshold value is constant:** set the threshold value of 100, and then change the different cluster radius. When the system cluster data using data mining algorithms, the results are as follows in theory:

Changing cluster radius will greatly affect the data clustering result. As the cluster radius reduces, the more network behavior pattern classes will be generated. Instead the cluster radius increases, the fewer network behavior pattern classes will be generated. The experimental result of affecting by clustering radius is shown in Table 3. In Table 3, when the threshold is a fixed value, changing the different clustering radius can be directly affected the number of clusters. When reduces the clustering radius, the system generated more network behavior pattern classes. When increases the clustering radius, the system generated fewer network behavioral pattern classes. The experimental result is the same as the theoretical analysis. The result also shows that the data which are calculated in the preliminary detection engine can be received in false detection rate Table 4.

From the Table 4, it shows that changing the cluster radius has a great effect on false detection rate of preliminary detection engine. When reducing the cluster radius, false detection rate becomes lower. When increasing cluster radius, false detection rate becomes larger.

False detection rate is an important parameter. The system must reduce the false detection rate and prevent discarding normal packet. The result of the experiment is the same as the theoretical analysis.

- **The case of cluster radius is constant:** set the clustering radius r = 5, and then change the different threshold. It can be obtained from theoretical analysis that the system will have more the network normal behavior patterns classes and fewer network abnormal behavior pattern classes when reducing the threshold. Instead, the system will have fewer the network normal behavior patterns classes and more network abnormal behavior pattern classes when increasing the threshold. The experimental result of influence of different threshold is shown in Table 5.

In Table 5, different thresholds have different results. As the threshold grows larger, the number of network normal behavior pattern class becomes fewer; the number of network abnormal behavior patterns class becomes more. The experimental result is consistent to the theoretical analysis.

Result also shows that different thresholds have a direct effect on the false detection rate, as shown in Table 6.

In Table 6, the larger threshold the system sets, the less false detection rate the system makes. It is the same as the theoretical analysis.

- **Results:** From the four tables (Table 3 to 6), the two important parameters (cluster radius and threshold) have a great influence on the clustering and false detection rate. When threshold is fixed, as the clustering radius increase, the network behavior pattern classes become fewer. When cluster radius is unchanged, as threshold value becomes lower, the false detection rate becomes higher. Therefore, according to the needs and actual situation of practical applications, we need to adjust cluster radius and threshold to achieve a satisfactory result.

CONCLUSION

Aiming at weakness of self-adaptation ability, low false alarm rate and high misinformation rate of the current most of the intrusion detection system, This study has designed and implemented an intrusion detection system framework based on data mining technology, and has introduced the process of correlation analysis data mining algorithm that how to construct into the intrusion detection model. The test results have shown that the intrusion detection based on data mining system, which overcomes certain limitations of the intrusion detection system, provides self-adaptability, improves the detection efficiency, and reduces the previous deviations caused by domain experts' hand writing mode.

ACKNOWLEDGMENT

The authors wish to thank the helpful comments and suggestions from my teachers and colleagues in intelligent detection and control lab at Zhengzhou. This study is financially supported by the National Natural Science Foundation of China (No. 61250007) and the National '863' Program of China (No. 2009AA012201).

REFERENCES

Chen, X., 2010. The approach of instruction detection based on data mining. J. Comput. Eng., 3(5): 156-161.

Gao, X. and M. Wang, 2003. Research intrusion detection system based on data mining technology. J. Northwestern Polytech. Univ., 21(4): 189-192.

Gu, J. and L. Sun, 2006. Application research of data mining technology for intrusion detection. J. Comput. Technol. Develop., 16(9): 243-246.

Jiang, J., H. Ma and D. Ren, 2000. A survey of intrusion detection research on network security. J. Softw., 11(11): 1460-1466.

Lu, Y., Y. Cao and J. Ling, 2003. The structure of intrusion detection system based on data mining approach. J. Wuhan Univ., 48(1): 63-66.

Ning, Y. and X. Guo, 2006. Data mining approaches for network intrusion detection. J. Comput. Measur. Control, 10(3): 189-192.

Su, H. and L. Fu, 2007. An intelligent model of instrusion detection based on data mining approach. J. Micro Comput. Inform., 23(3): 74-77.

Xu, S. and L. Cai, 2008. Intrusion detection system based on data mining technology. J. Elect. Des. Eng., 17(8): 3-5.